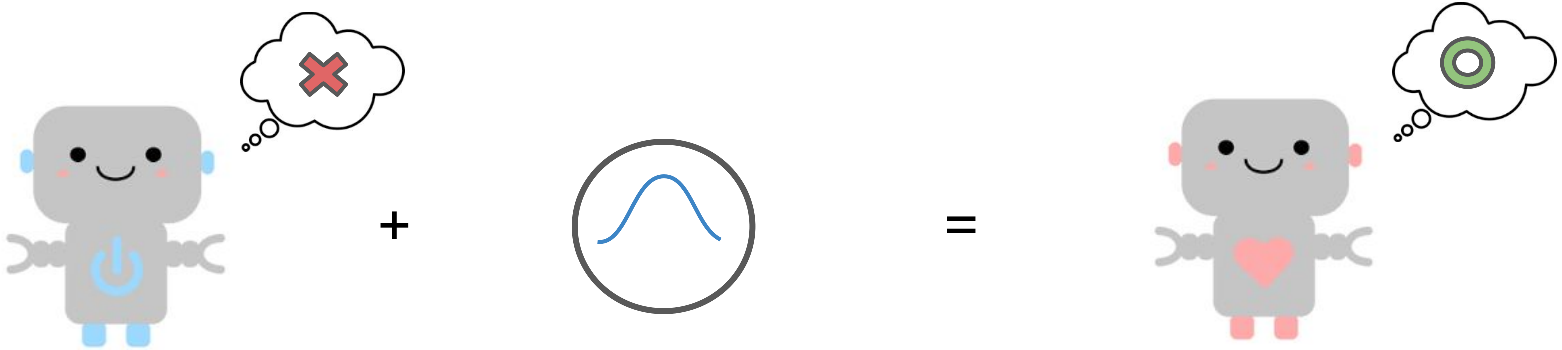
Meta-Amortized Variational Inference and Learning

Kristy Choi

CS236: December 4th, 2019

Probabilistic Inference

Probabilistic inference is a particular way of viewing the world:

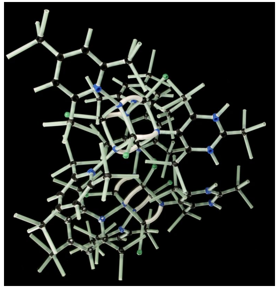


Typically the beliefs are “hidden” (unobserved), and we want to model them using latent variables.

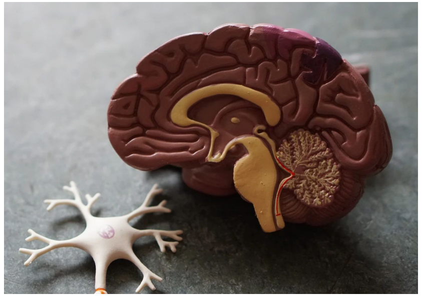
Probabilistic Inference

Many machine learning applications can be cast as probabilistic inference queries:

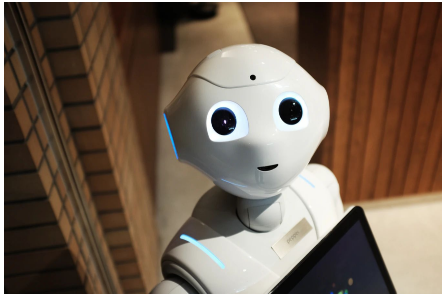
Medical diagnosis



Bioinformatics



Human cognition



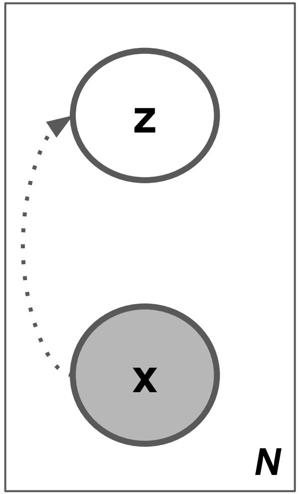
Computer vision

Medical Diagnosis Example

observed symptoms $\mathbf{x} \in \mathcal{X}$

identity of disease $\mathbf{z} \in \mathcal{Z}$

Goal: Infer identity of disease given a set of observed symptoms from a patient population.



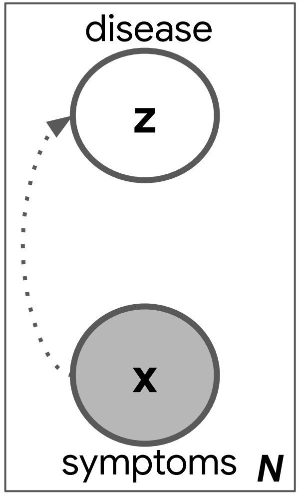
Exact Inference

intractable integral $\int_z p(\mathbf{x},\mathbf{z})dz$

$$\boxed{p(\mathbf{z}|\mathbf{x})} = p(\mathbf{x},\mathbf{z}) / \boxed{p(\mathbf{x})}$$

$$\approx$$

family of tractable distributions $q_\psi \in \mathcal{Q}$



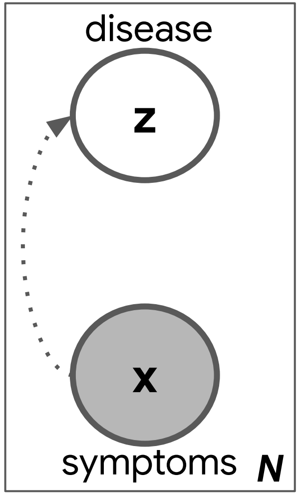
Marginal is intractable, we can't compute this even if we want to

Approximate Variational Inference

dependence on **x**: learn new q per data point

$$\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\underbrace{\left[\max_{\psi} \mathbb{E}_{\boxed{q_{\psi}(\mathbf{z})}} \log \frac{p(\mathbf{x},\mathbf{z})}{q_{\psi}(\mathbf{z})}\right]}$$

→ turned an intractable inference problem into an optimization problem

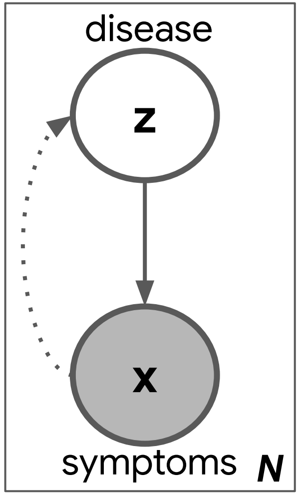


Amortized Variational Inference

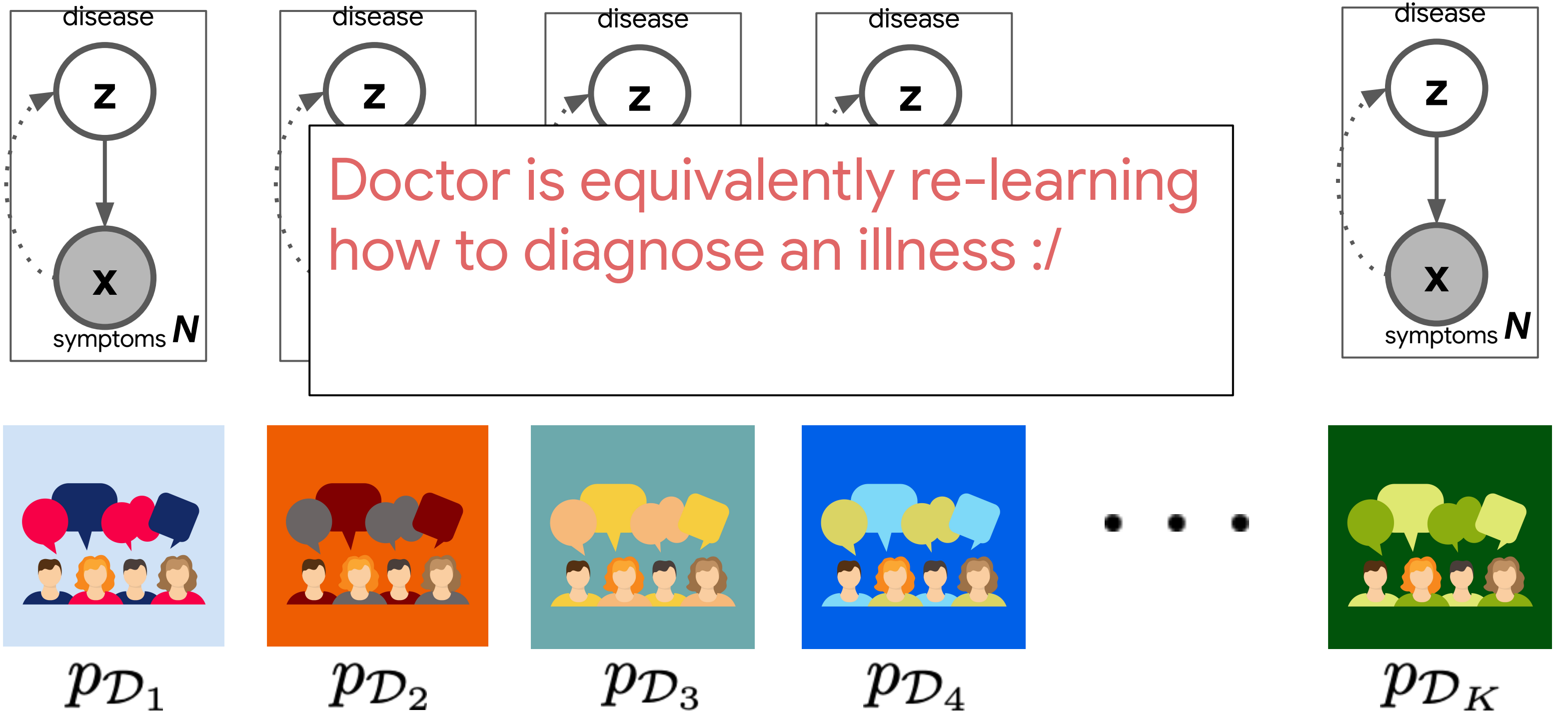
deterministic mapping predicts **z** as a function of **x**

$$\max_{\phi} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\underbrace{\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}\log\frac{p_{\theta}(\mathbf{x},\mathbf{z})}{q_{\phi}(\mathbf{z}|\mathbf{x})}}\right]$$

→ scalability: VAE formulation



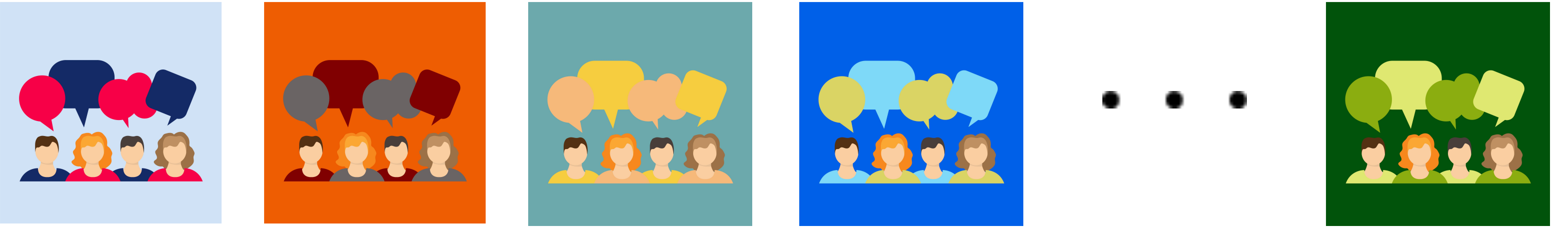
Multiple Patient Populations



Doctor is equivalently re-learning how to diagnose an illness :/

Multiple Patient Populations

Share statistical strength across different populations to infer latent representations that transfer to similar, but previously unseen populations (distributions)



$p_{\mathcal{D}_1}$ $p_{\mathcal{D}_2}$ $p_{\mathcal{D}_3}$ $p_{\mathcal{D}_4}$ $\cdots$ $p_{\mathcal{D}_K}$

(Naive) Meta-Amortized Variational Inference

$$\mathbb{E}_{p_{\mathcal{D}_i}\sim p_{\mathcal{M}}}\left[\boxed{\max_{\phi}}\,\mathbb{E}_{p_{\mathcal{D}_i}(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log\frac{p_{\theta_i}(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right]$$

$p_{\mathcal{D}_1}$

$p_{\mathcal{D}_2}$

$p_{\mathcal{D}_3}$

$\cdots$

$p_{\mathcal{D}_K}$

$\sim p_{\mathcal{M}}$

meta-distribution

Meta-Amortized Variational Inference

$$\max_{\phi} \mathbb{E}_{p_{\mathcal{D}_i} \sim p_{\mathcal{M}}} \left[\mathbb{E}_{p_{\mathcal{D}_i}(\mathbf{x})} \left[\mathbb{E}_{\boxed{g_\phi(p_{\mathcal{D}_i}, \mathbf{x})}} \log \frac{p_{\theta_i}(\mathbf{x}, \mathbf{z})}{g_\phi(p_{\mathcal{D}_i}, \mathbf{x})(\mathbf{z})} \right] \right]$$

shared meta-inference network

$p_{\mathcal{D}_1}$ $p_{\mathcal{D}_2}$ $p_{\mathcal{D}_3}$ $\cdots$ $p_{\mathcal{D}_K}$

$\sim p_{\mathcal{M}}$

meta-distribution

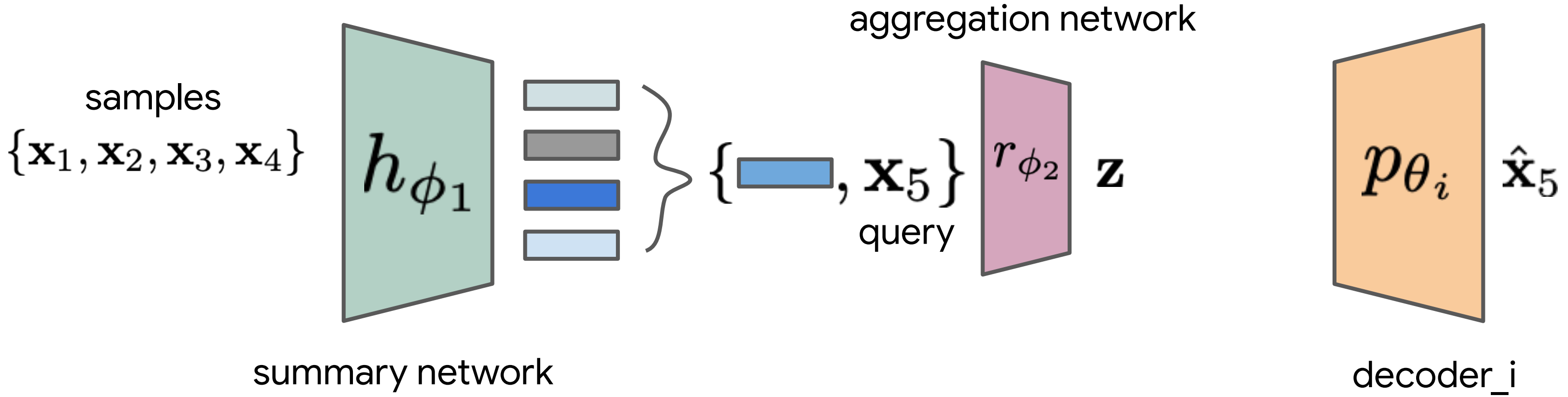
Meta-Inference Network

- Meta-inference model $g_\phi(p_{\mathcal{D}_i}, \mathbf{x})(\mathbf{z})$ takes in 2 inputs:
    - Marginal $p_{\mathcal{D}_i}$
    - Query point $\mathbf{x}$
- Mapping $g_\phi : \mathcal{M} \times \mathcal{X} \to \mathcal{Q}$
- Parameterize encoder with neural network
- Dataset $\mathcal{D}_i$: represent each marginal distribution as a set of samples

$$\mathcal{D}_i = \{\mathbf{x}_j \sim p_{\mathcal{D}_i}(\mathbf{x})\}_{j=1}^{N}$$

In Practice: MetaVAE

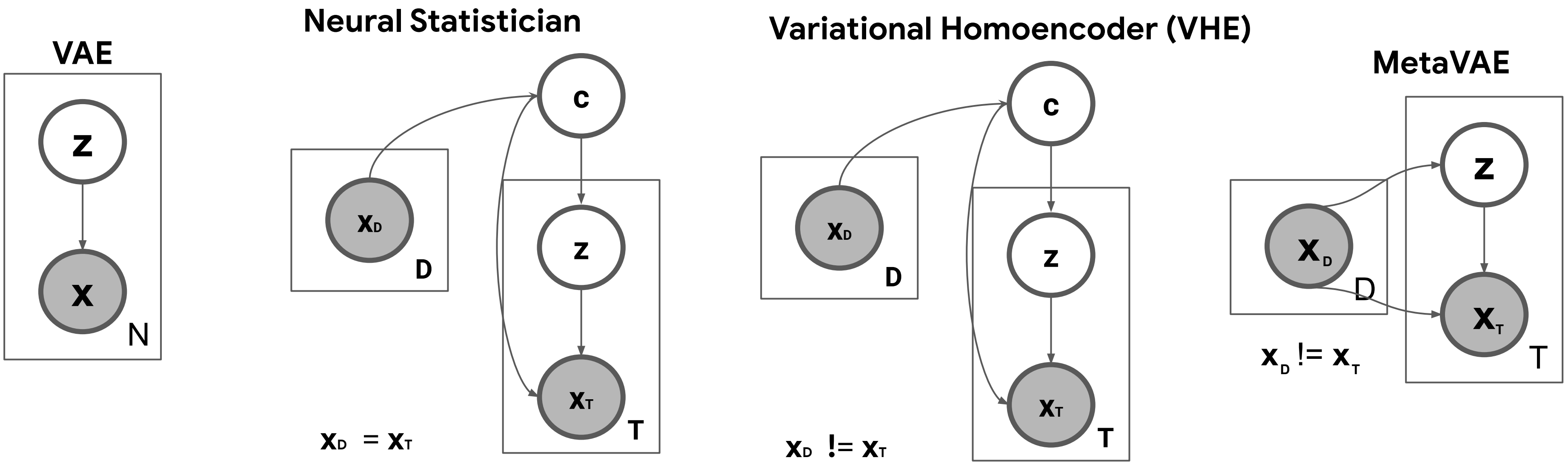
$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \sim$  $p_{\mathcal{D}_i}$



Summary network ingests samples from each dataset

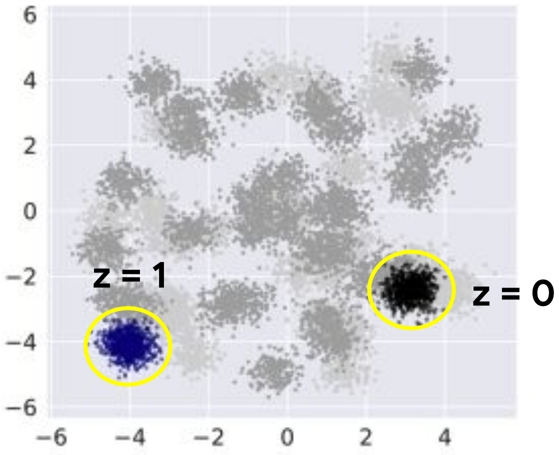
Aggregation network performs inference $\phi = \{\phi_1, \phi_2\}$

Related Work



Avoid restrictive assumption on global prior over datasets p(c)

Intuition: Clustering Mixtures of Gaussians

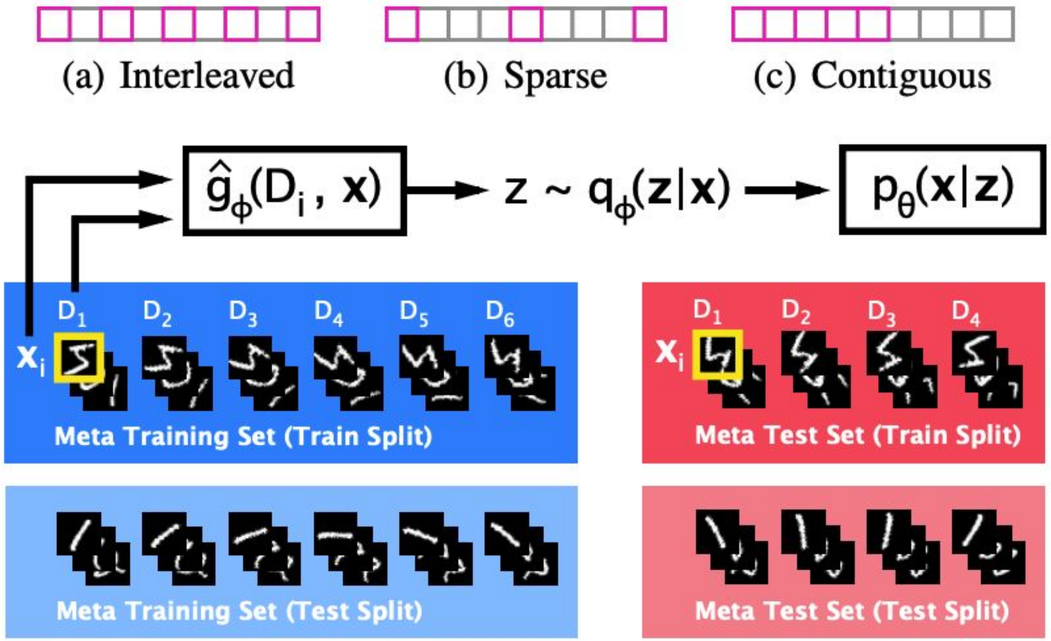


$$\mu_1, \mu_2 \sim \mathcal{U}(-5, 5)$$

$$p_{\mathcal{D}_i}(\mathbf{x}) = \frac{1}{2}\mathcal{N}(\mu_1, 0.1) + \frac{1}{2}\mathcal{N}(\mu_2, 0.1)$$

Learns **how to cluster**: for 50 datasets, MetaVAE achieves 9.9% clustering error, while VAE gets 27.9%

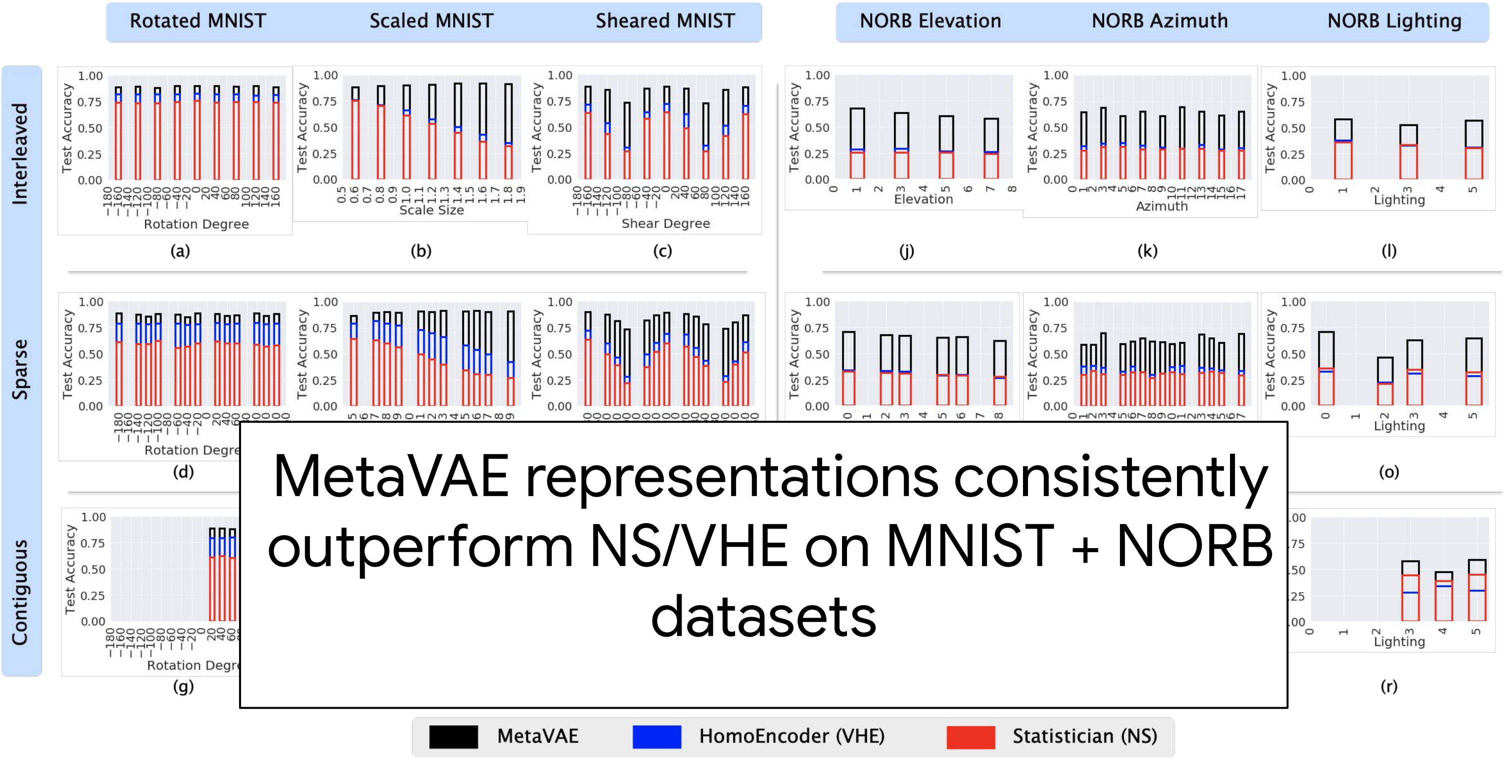
Learning Invariant Representations



(d) Meta-Inference Pipeline

- Apply various transformations
- Amortize over subsets of transformations, learn representations
- Test representations on held-out transformations (classification)

Invariance Experiment Results



MetaVAE representations consistently outperform NS/VHE on MNIST + NORB datasets

Analysis

Model Dataset	Rotation	Scale	Skew
Rotated MNIST	**1.65**	4.44	4.09
Scaled MNIST	5.44	**2.16**	4.92
Skewed MNIST	3.79	4.89	**1.47**
Model Dataset	**Elevation**	**Azimuth**	**Lighting**
NORB Elevation	**0.39**	1.16	1.27
NORB Azimuth	1.42	**0.44**	1.26

MetaVAE representations tend not to change very much within a family of transformations that it was amortized over, as desired.

Conclusion

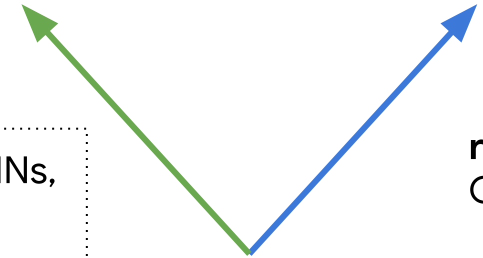
- Limitations
  - No sampling
  - Semi-parametric
  - Arbitrary dataset construction
- Developed an algorithm for a family of probabilistic models: meta-amortized inference paradigm
- MetaVAE learns transferrable representations that generalize well across similar data distributions in downstream tasks
- Paper: https://arxiv.org/pdf/1902.01950.pdf

Encoding Musical Style with Transformer Autoencoders

Generative Models for Music

- Generating music is a challenging problem, as music contains structure at multiple timescales.
  - Periodicity, repetition
- Coherence in style and rhythm across (long) time periods!

symbolic: RNNs, LSTMs, etc.

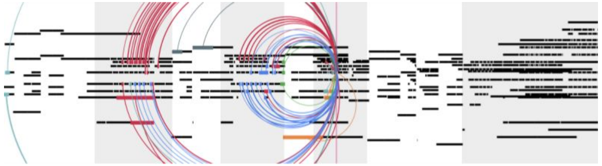


raw audio: WaveNet, GANs, etc.

Music Transformer

- Symbolic: event-based representation that allows for generation of expressive performances (without generating a score)
- Current SOTA in music generation
  - Can generate music over 60 seconds in length
- Attention-based
  - Replaces self-attention with relative attention

What We Want



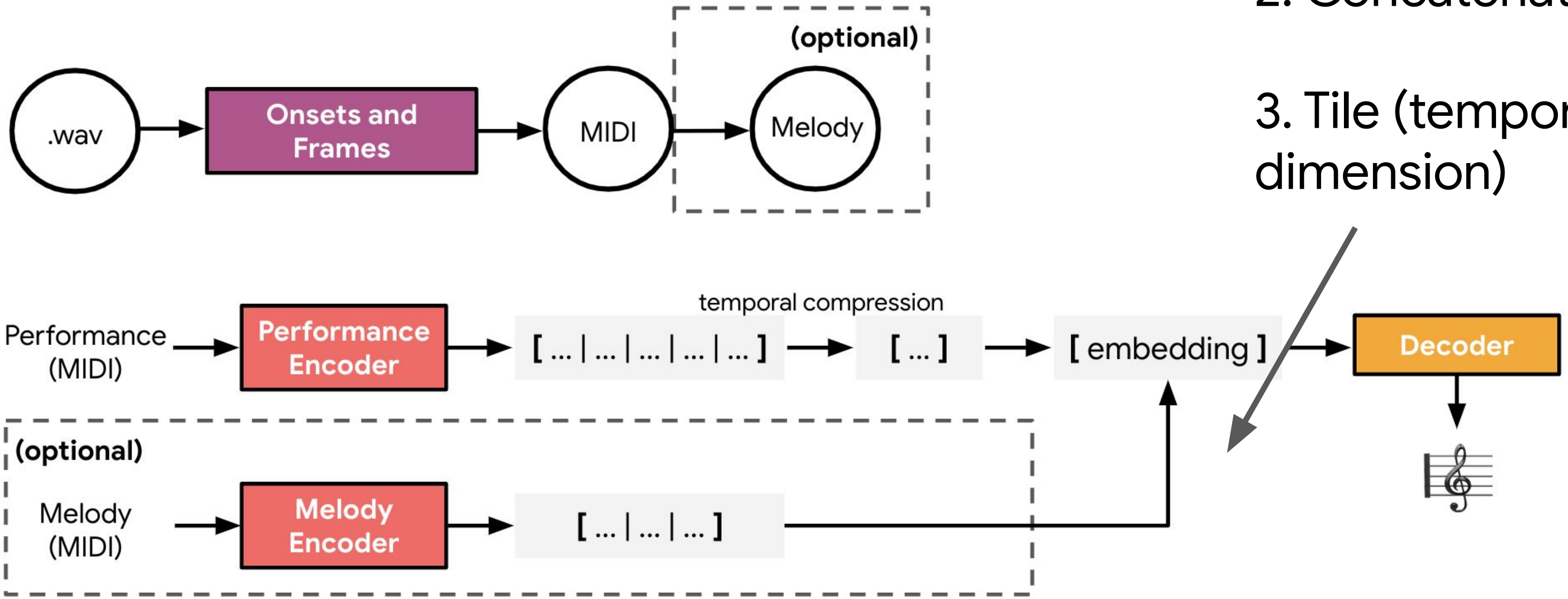
- Control music generation using either (1) performance or (2) melody + perf as conditioning
- Generate pieces that sound similar in style to input pieces!

Transformer Autoencoder

1. Sum

2. Concatenation

3. Tile (temporal dimension)



Quantitative Metrics

MAESTRO	ND	PR	MP	VP	MV	VV	MD	VD	Avg
Melody & perf. (ours)	**0.650**	**0.696**	0.634	0.689	**0.692**	0.732	**0.582**	**0.692**	**0.67**
Perf-only (ours)	0.600	0.695	**0.657**	**0.721**	0.664	**0.740**	0.527	0.648	0.66
Melody-only	0.609	0.693	0.640	0.693	0.582	0.711	0.569	0.636	0.64
Unconditional	0.376	0.461	0.423	0.480	0.384	0.588	0.347	0.520	0.48
Internal Dataset									
Melody & perf (ours)	**0.646**	**0.708**	0.610	0.717	**0.590**	**0.706**	**0.658**	**0.743**	**0.67**
Perf-only (ours)	0.624	0.646	0.624	0.638	0.422	0.595	0.601	0.702	0.61
Melody-only	0.575	0.707	**0.662**	**0.718**	0.583	0.702	0.634	0.707	0.66
Unconditional	0.476	0.580	0.541	0.594	0.400	0.585	0.522	0.623	0.54

Table 4: Average ... different condi-
tioning. Unconditi... The metrics are
described in detail ...d for the listener
study shown in the

Transformer autoencoder (both performance-only and melody & perf) outperform baselines in generating similar pieces!

Samples

Twinkle, Twinkle melody

Conditioning Performance

Generated Performance: “Twinkle, Twinkle” in the style of the above performance

Claire de Lune

Conditioning Performance

Generated Performance: “Claire de Lune” in the style of the above performance

Conclusion

- Developed a method for controllable generation with high-level controls for music
  - Demonstrated efficacy both quantitatively and through qualitative listening tests
- Thanks!
  - **Stanford:** Mike Wu, Noah Goodman, Stefano Ermon
  - **Magenta @ Google Brain:** Jesse Engel, Ian Simon, Curtis “Fjord” Hawthorne, Monica Dinculescu

References

1. Edwards, H., and Storkey, A. Towards a neural statistician. 2016
2. Hewitt, L. B., Nye, M. I.; Gane, A.; Jaakkola, T., and Tenenbaum, J.B. Variational Homoencoder. 2018
3. Kingma, D.P., and Welling, M. Auto-encoding variational bayes. 2013
4. Gershman, S., and Goodman, N. Amortized inference in probabilistic reasoning. 2014
5. Jordan, M. I.; Ghahramani, Z.; Jaakkola, T.S.; and Saul, L.K. An introduction to variational methods for graphical models. 1999
6. Blei, D. M.; Kuckelbir, A.; and McAuliffe, J.D. Variational inference: a review for statisticians. 2017
7. Huang, C.Z.; Vaswani, A., Uskoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., Eck, D. Music Transformer. 2019
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention is all you need. 2017
9. Shaw, P., Uszkoreit, J., Vaswani, A. Self-Attention with relative position representations. 2018
10. https://magenta.tensorflow.org/music-transformer
11. Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., Roberts, A. Adversarial Neural Audio Synthesis. 2019
12. Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. 2016
13. Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., Kavukcuoglu, K. Efficient Neural Audio Synthesis. 2018